融合多源网络评估数据及URL特征的钓鱼网站识别技术研究*

胡忠义　王超群　吴　江
(武汉大学信息管理学院　武汉 430072)
(武汉大学电子商务研究与发展中心　武汉 430072)

摘要:【目的】充分利用多源网络评估数据和 URL 异常特征数据，研究提高钓鱼网站识别准确性的可行性方案。【方法】采用 8 种机器学习技术，对比研究网络评估数据与传统的 URL 异常特征数据在钓鱼网站识别中的性能，并融合两类数据研究进一步提高钓鱼网站识别准确性的可行性方案。【结果】在钓鱼网站识别中，相比于传统的 URL 异常特征，利用网络评估数据可以取得更好的识别效果。融合两类数据对于提高识别准确性有一定帮助。【局限】未考虑钓鱼网站与正常网站的数量存在严重的不均衡问题。【结论】充分利用多源网络评估数据和 URL 异常特征数据识别钓鱼网站的方法是比较合理和有效的，对后续相关研究具有一定的借鉴意义。

关键词: 数据挖掘　钓鱼网站识别　机器学习
分类号: G353

1　引　言

近年来，随着互联网的高速发展，网民数量急剧增加，在电子商务及电子金融产业日渐繁荣的同时，网络安全问题日益凸显。病毒、盗号、木马、钓鱼等黑客行为对互联网环境造成极其恶劣的影响，其中钓鱼网站的危害尤其严重。钓鱼网站是一种基于社会工程学的攻击手段，不法分子通过垃圾邮件、聊天工具、手机短信或网页虚假广告发送大量声称来自于银行或其他知名机构的欺骗性信息，意图引诱用户给出敏感信息(如用户名、口令、手机号、银行账号和密码)。根据 2016 年 6 月中国互联网协会发布的《中国网民权益保护调查报告(2016)》，中国互联网用户近一年来因个人信息泄露、诈骗信息等问题，导致总体损失约 915 亿元。

为了阻止用户访问钓鱼网站并最大程度地减小用户损失，互联网厂商采用黑白名单技术，推出了用于识别钓鱼网站的浏览器检测插件[1-2]。但是，随着钓鱼网站数量的急剧增加，黑白名单技术并不能有效地解决钓鱼网站的识别问题，不少学者尝试基于 URL 异常特征构建识别模型，用以有效识别钓鱼网站[3-9]。有学者借助互联网上可公开获取的网站评估数据，开展了基于网络评估数据的钓鱼网站识别技术研究[10]。尽管基于 URL 异常特征的识别技术和基于网络评估数据的识别技术均取得了较好的效果，但尚未有研究对比这两类技术在钓鱼网站识别中的性能。因此，本文采用 8 种机器学习技术，对比两类识别技术在钓鱼网站识别中的性能，并尝试融合两类多源数据，研究进一步提高钓鱼网站识别准确率的可行性方案。

2　相关研究

鉴于钓鱼网站危害极大，国内外研究提出了多种不同的解决方案，如基于黑名单的识别技术和基于 URL 异常特征的机器学习识别技术。其中，基于黑白名单的识别技术多采用浏览器插件形式实现[1-2]，如微

通讯作者: 胡忠义, ORCID: 0000-0002-1113-0199, E-mail: zhongyi.hu@whu.edu.cn。

*本文系国家自然科学基金面上项目“创新 2.0 超网络中知识流动和群集交互的协同研究”(项目编号: 71373194)和国家自然科学基金青年基金项目“基于集成学习的区间型电力负荷预测技术研究”(项目编号: 71601147)的研究成果之一。

软IE浏览器的Phishing Filter、谷歌出品的Google Safe Browsing、搜狗网页安全卫士等。基于事先维护的钓鱼网站黑名单列表，当用户访问的网站在黑名单中时，浏览器会弹出警示框，提醒用户当前访问的网站是钓鱼网站，进而阻止用户访问该钓鱼网站。虽然这些基于黑名单技术的检测方式简单直接，但在实践运行中效果欠佳。如，为了测试现有浏览器厂商或第三方厂商提供的各种浏览器防钓鱼网站插件，Zhang 等[2]设计了一种自动检测平台。针对该平台收集的 200 个钓鱼网址和 516 个合法网址，10 个流行的防钓鱼网站插件中只有两个工具可以识别出 60%以上的钓鱼网址。究其原因，主要是由于黑名单往往是通过网民举报和人工审核等方式建立的，具有一定的滞后性；同时，随着钓鱼网站数量的急剧增加，使得建立一份完整的黑白名单的难度越来越大。因此这种方法虽然技术简单，但无法从本质上检测钓鱼网络攻击。

基于 URL 异常特征的识别技术利用钓鱼网站的 URL 特征，基于机器学习算法构造用以识别钓鱼网站的分类器模型。如，Blum 等[3]从 URL 中提取词汇特征并构建可信度加权的分类算法。黄华军等[4]通过分析网站域名结构上的特征和语义上的特征，抽取 10 多个有效特征，用以构建和测试基于支持向量机的分类模型，达到了较好的识别效果。Ma 等[5-6]构建的恶意网站的识别模型中，则采用了 URL 词汇特征和主机特征。基于 URL 中提取出的敏感特征，曾传璜等[7]设计了改进的 AdaCostBoost 算法，实验结果表明，该检测方法具有较优的检测性能。相比黑白名单识别技术，基于 URL 异常特征的钓鱼网站识别技术不再需要人工去标注钓鱼网站，工作效率有了很大提高[8]，且能够在一定程度上应对钓鱼网站的快速变化。但是，URL仿照性较强，仅仅通过 URL 异常特征识别钓鱼网站可能会造成较高的误判率和漏判率[9]，风险较大。

近年，Hu 等[10]融合网站的多源网络评估数据构建多种机器学习模型，用于识别钓鱼网站。该研究利用互联网上可公开获取的评估数据(如知名互联网公司测评的域名评估数据、社交平台关注数据等)，构造网站评估数据的特征向量，并采用多种稳健的机器学习算法，构建钓鱼网站识别模型。结果表明各机器学习模型可以较好地利用网络评估数据识别钓鱼网站。该方法符合当前大数据分析中充分融合多源数据的潮流，实现识别钓鱼网站的目的。然而，该研究仅仅验证了各机器学习模型利用网络评估数据在识别钓鱼网站中的有效性，而所提方法与传统方法如基于URL异常特征的识别技术相比，是否具有更好的识别效果，并未加以验证，且进一步融合两类特征变量是否能更好地提高钓鱼网站的识别性能也仍未知。

因此，本文针对钓鱼网站识别问题，对比研究多源网络评估数据与 URL 特征数据在钓鱼网站识别中的性能，并融合两类多源数据特征，进一步探究提高钓鱼网站识别准确性的可行性方案，以期为相关的钓鱼网站识别研究提供参考。

本文的主要创新包括：采用 8 种经典的机器学习技术系统地评价基于不同特征变量识别钓鱼网站的性能；采用 Boruta 技术进行特征选择，剔除冗余特征变量，提高模型性能；基于多个评价指标，全面对比分析了 URL 特征数据、多源网络评估数据及融合两类特征数据识别钓鱼网站的性能。

3 钓鱼网站识别模型

为了更为准确地识别钓鱼网站，本文提出融合多源网络评估数据及 URL 特征的识别模型。图 1 描述了该技术的详细流程。主要包括数据采集与预处理、特征选择、模型构建与验证三部分。

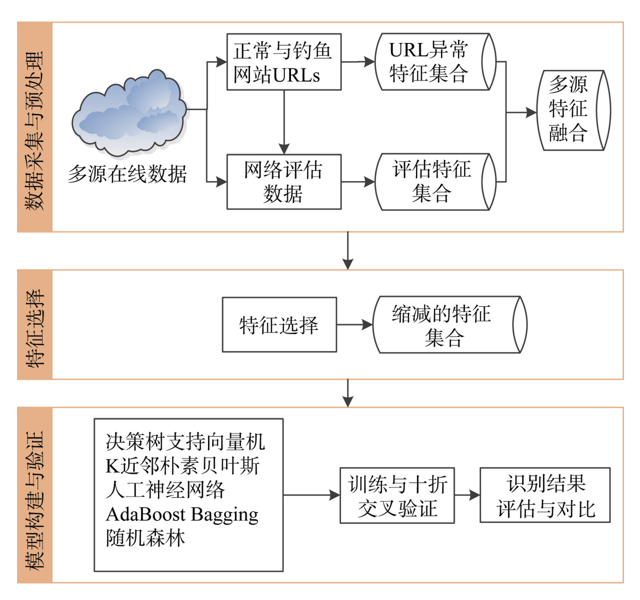


图 1 识别技术流程

(1) 数据采集与预处理

为了构建融合多源网络评估数据及 URL 特征的

钓鱼网站识别模型，从 PhishTank、Alexa 等钓鱼网站名录和知名站点名录收集网站 URLs 集；从 Moz、Majestic、Google、Alexa 等第三方知名网站评估平台收集网络评估数据；经数据清理、筛选、抽取等处理后得到 URL 异常特征变量和网络评估特征变量，并融合两类特征向量得到多源特征集合。

(2) 特征选择

上述收集提取的 URL 异常特征、网络评估特征及融合两类特征的多源特征集合，可能存在一些不相关或者冗余的特征变量，这些变量不仅会影响模型的识别精度，还会增加模型的复杂度，进而降低效率，因此使用特征选择技术去除多余变量是必要的。为此，本研究采用 Boruta 算法[11]进行特征变量的筛选，以去除多余的变量、提高模型识别精度。

Boruta 是一种特征选择算法，通过创建混合副本的所有特征增加了随机性，而且使用特征重要性指标(默认设定为平均减少精度)评估每个特征的重要性，每次迭代的时候都会比较每一个真实的特征变量是否比最好的阴影特征具有更高的重要性，以此为依据删除不重要的特征[11]。在 Boruta 执行完变量筛选后，会对数据集中变量的意义给出明确的解释。

(3) 模型构建与验证

为了有效评估不同特征变量在钓鱼网站识别中的性能，本文采用决策树、支持向量机、K 近邻法、朴素贝叶斯、人工神经网络、AdaBoost、Bagging、随机森林 8 种经典的机器学习技术构建识别模型，这些模型在数据挖掘和机器学习领域得到广泛应用，同时也是钓鱼网站识别研究中的常用技术。其中，决策树、支持向量机、K 近邻法、朴素贝叶斯、人工神经网络属于单一模型，AdaBoost、Bagging、随机森林属于集成模型。值得注意的是，集成模型集合了多个弱学习器，相比于单一模型，往往具有更高的准确性[12-13]。基于以上 8 种机器学习技术，采用十折交叉验证，通过准确率(Accuracy)、查全率(Recall)、查准率(Precision)、F 值(F-measure) 4 种评价指标，全面评估和对比基于 URL 异常特征、网络评估特征、多源融合特征的钓鱼网站识别模型和性能。

4　数据收集与处理

4.1　数据收集

为了研究网络评估数据与传统的 URL 异常特征数据在识别钓鱼网站中的性能，本研究既获取了网站的 URL 特征数据，又获取了网站的网络评估数据。一共获取了 2 000 条 URL 数据，这些 URL 截至 2016 年 7 月 31 日仍然可以解析。为了消除数据不平衡问题，数据集中合法网址和钓鱼网址各占一半。1 000 条钓鱼网站数据从 PhishTank①中获取。1 000 条合法网站数据从 Alexa②获取，且这些合法网站既有访问量很大的网站，也有访问量极少的，并在 Alexa 中排名特别靠后(如 1 000 万以后)的网站。

4.2　数据描述

(1) URL 特征变量

通过对已有相关文献的分析，选取 7 个钓鱼网站的特征变量，组成 URL 特征向量 FV:

FV=<F1, F2, F3, F4, F5, F6, F7>

F1: length，URL 长度，一般可信网站 URL 的长度小于 23，URL 过长的网站，就有可能是钓鱼网站。

F2: isContainIp，URL 中是否含有 IP 地址，为逃避域名的注册或用户的检查，用十进制掩饰的基于 IP 地址的 URL 地址是一种在钓鱼网站中常用的手段。

F3: isContainSensitiveWord，URL 中是否包含敏感词汇，敏感词汇包括 admin、login、manage、root、account、bank、password 等，当网址中出现较多敏感词汇时，可能就是钓鱼网站为了获取用户的信息而设置的。

F4: isContainSpecailCharactor，URL 中是否出现异常字符，异常字符包括-、~、!、@、#、$、%、*等，如果网址的异常字符过多，该网站很有可能就是钓鱼网站。

F5: countOfDot，URL 域名级数，当 URL 中包含过多的域名级数时，很可能是钓鱼网站模仿合法网站，故意加入产品信息。

F6: countOfSlash，URL 目录级数，设置较多的路径级数时可以让用户眼花缭乱以至于无法辨别出是钓鱼网站。

F7: count，URL 中长单词(长度超过 20)的个数，正常网站中出现长单词的次数很少。

(2) 互联网评估数据

针对互联网评估数据，分别从 Moz、Majestic、Google、Alexa 共4家知名网络采集多源网站评估数据，

①http://www.phishtank.com/.

②http://www.alexa.com/.

经处理，得到包含16个变量的评估数据特征向量FV:

FV=<F8, F9, F10, F11, F12, F13, F14, F15, F16, F17, F18, F19, F20, F21, F22, F23>

①Moz评估数据

F8: Moz's Domain Authority, Moz公司给出的域名在搜索引擎中排名的预测。

F9: MozRank, 代表一个链接流行度评分。

F10: Moz's Total Backlinks, 反映一个网站的所有反向链接，反向链接越多，说明这个网站越受欢迎。

②Majestic评估数据

F11: Majestic's Citation Flow, 用来度量引用来源，通过引用排名，显示一个网站的受欢迎程度。

F12: Majestic's Trust Flow, 用来度量信任来源，表明一个网站和可信赖网站的亲密程度。

F13: Majestic's Backlinks, 反映网站反向链接的指标。

F14: Majestic's Reference Domains, 引用域，是指外部链接指向当前网站的个数。

社交网站分享度：可以反映出各个网站在社交网站的受欢迎程度，社交网站的受欢迎度排名越高，网站的权威性越高。

F15: Facebook Shares, 在Facebook的受欢迎程度。

F16: Twitter Tweets, 在Twitter的受欢迎程度。

F17: Google Plus Shares, 在Google Plus的受欢迎程度。

③Google评估数据

F18: Google's Page Rank, 是Google通过网站之间的超链接关系确定的网站排行榜。

F19: Google's Page Speed, 是Google评估网页加载速度的指标。

④Alexa评估数据

F20: Alexa's Rank, 通过网站的访问量确定网站排名，访问量越大，排名越靠前，网站越受欢迎。

F21: Alexa's 1 Month Reach, 网站最近1个月的平均每天访问量。

F22: Alexa's 3 Month Reach, 网站最近3个月的平均每天访问量。

F23: Alexa's Median Load, 使用Alexa特有的算法计算出的页面平均加载速度。

4.3 评估方式

判断一个网站是钓鱼网站，还是正常网站，是典型的二分类问题。在现实生活中正常网站的数量远多于钓鱼网站的数量，钓鱼网站更容易出现错分，另外钓鱼网站的错分代价更大，因此钓鱼网站的识别率更重要。所以本文不采用总体分类性能指标，而是采用二分类问题的混合矩阵进行评估，如表1所示。

表1 二分类的混合矩阵

	判断是正常网站	判断是钓鱼网站
实际是正常网站	TN	FP
实际是钓鱼网站	FN	TP

其中，钓鱼网站样本为P，正常网站样本为N，FP是指将正常网站样本错分成钓鱼网站的数目，FN是指将钓鱼网站样本错分成正常网站的数目，TP和TN分别表示钓鱼网站和正常网站样本被正确分类的数目。

据此得到4类性能评价指标，分别如公式(1)–公式(4)所示。

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$\text{Recall}=\frac{TP}{TP+FN} \tag{2}$$

$$\text{Precision}=\frac{TP}{FP+TP} \tag{3}$$

$$\text{F-measure}=\frac{2\times \text{Recall}\times \text{Precision}}{\text{Recall}+\text{Precision}} \tag{4}$$

从公式(4)可知，性能指标F-measure既考虑钓鱼网站样本的查全率又考虑查准率，只有在查全率和查准率的值都比较大的前提下，F-measure值才会很大，因此能综合地体现出分类器对正常网站和钓鱼网站的分类效果，而且侧重于体现钓鱼网站样本的分类效果。

5 实验结果和分析

为了探究提高钓鱼网站识别准确率的可行性方案，对比研究基于URL异常特征的识别技术、基于网络评估数据的识别技术和融合两类特征变量的识别技术在钓鱼网站识别中的性能表现。为了消除指标之间的量纲影响，对特征向量进行归一化处理，随后采用8种机器学习技术分别对三类特征向量构建识别模型，并最终通过指标对比分析基于三类不同特征向量的识别模型的识别性能。

数据采集、处理及识别模型的训练等所有实验均在R语言环境下进行，实验涉及到的"rpart"、"e1071"、"kknn"、"nnet"、"adabag"、"randomForest"、"Boruta"等程序包均可下载[①]。

①程序包下载地址: http://cran.r-project.org/.

5.1 基于 URL 异常特征的识别

(1) 特征选择

在收集的 URL 特征中，可能存在一些不相关或者冗余的变量，这些变量不仅会影响模型识别精度，还会增加模型的复杂度，进而降低效率。因此，首先基于 Boruta 特征选择方法[11]，对含有 7 个 URL 特征变量的数据进行变量筛选，结果如图 2 所示。

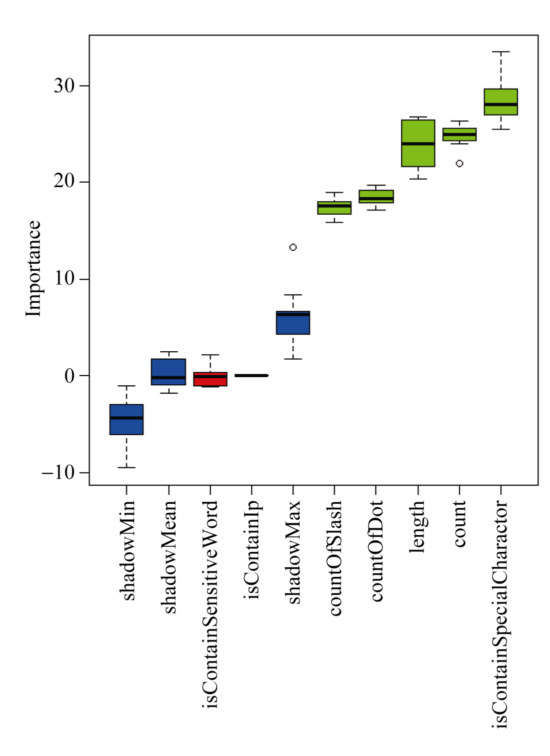


图 2 基于 Boruta 的 URL 特征选择

图 2 展示了 Boruta 计算的各变量的重要性，其中，红色和绿色的盒状图分别代表拒绝变量和确认变量的 Z 分数。蓝色的盒状图对应一个阴影变量的最小、平均和最大 Z 分数。可知，isContainIp(是否含有 IP)和 isContainSensitiveWord(是否含有敏感词汇)两个变量被拒绝，其余的 5 个被确认。

(2) 结果分析

基于已确认的 5 个 URL 特征变量，采用决策树、支持向量机、K 近邻法、朴素贝叶斯、人工神经网络、AdaBoost、Bagging、随机森林 8 种机器学习技术分别构建识别模型，且每个模型均采用十折交叉验证的方式进行训练和测试，每组实验均重复 10 次以防止随机影响，最后计算这些实验的各评测指标的统计平均值。实验结果如表 2 所示。

表 2 8 种方法的评估结果

方法	准确率	查准率	查全率	F 值
决策树	0.5935	0.9099	0.2150	0.3433
SVM	0.6340	0.7744	0.3780	0.5074
K 近邻法	0.6205	0.6411	**0.5610**	**0.5954**
朴素贝叶斯	0.5990	0.9720	0.2040	0.3362
人工神经网络	0.6420	0.7535	0.4290	0.5457
AdaBoost	0.6435	0.7500	0.4400	0.5534
Bagging	**0.6445**	0.7587	0.4260	0.5443
随机森林	0.6390	0.7828	0.3850	0.5155

从表 2 可以看出，仅仅使用 URL 异常特征进行网站识别的性能不是很好。其中，F 值最高的为 K 近邻法(0.5954)，且该方法的查全率 0.5610 也是所有方法中最高的，但准确率和查准率相比三种集成模型略差一些。决策树、SVM、朴素贝叶斯三种单一方法虽然有相对较高的查准率，但是其查全率和 F 值很差，这说明它们将大多数钓鱼网站都识别为正常网站。三种集成模型(AdaBoost、Bagging 和随机森林)相对其他模型，4 个性能指标都比较适中，表现比较稳健，这主要在于这些模型是由众多弱模型集成而来，受噪声等随机因素的影响相对个体模型而言比较小。

总体来看，仅仅使用 URL 异常特征进行钓鱼网站的识别，效果不是很好，这主要是由于 URL 样式极易模仿和学习，导致 URL 异常特征特别有限，因而仅仅依赖 URL 异常特征进行钓鱼网站的识别是远远不够的。

5.2 基于网络评估数据的识别

(1) 特征选择

采用 Boruta 进行特征变量的筛选，以剔除冗余变量。各变量的重要性和检测结果如图 3 所示。

可知，16 个变量均被认为是重要的。其中，GooglePageRank、RefDomains 和 GooglePlusShares 是所有变量中最重要的三个。

(2) 结果分析

基于筛选得到的 16 个网络评估变量构建识别模型，实验结果如表 3 所示。

从表 3 可以看出，使用网络评估数据进行网站识别的准确率较高，除了朴素贝叶斯算法外，其余算法的准确率都在 0.85 以上，而且 F 值在 0.88 以上，相比仅利用 URL 异常特征进行识别有很大的提升。其中，

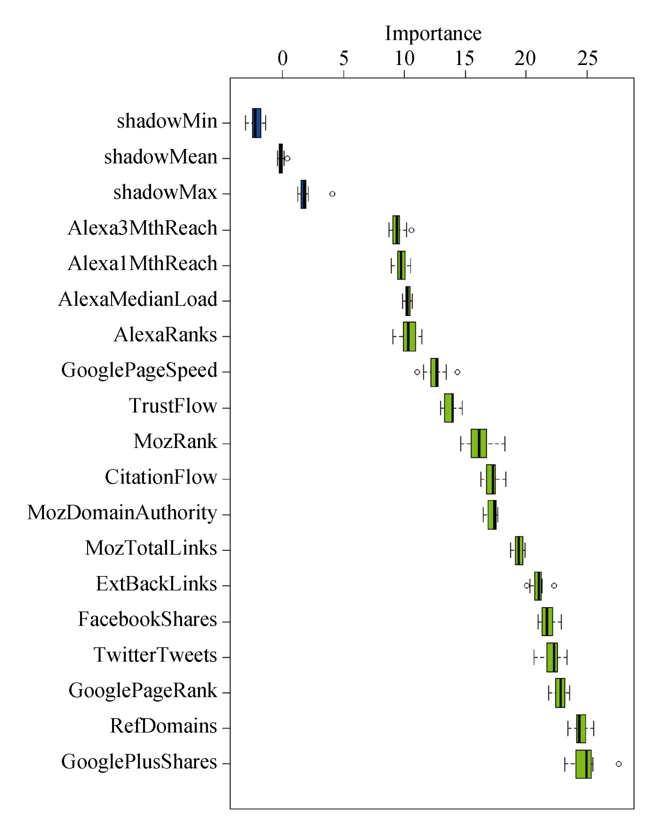


图 3 基于 Boruta 的网络评估数据特征选择

表 3 8 种方法的评估结果

方法	准确率	查准率	查全率	F 值
决策树	0.8810	0.8576	0.9160	0.8845
SVM	0.9145	0.9026	0.9310	0.9159
K 近邻法	0.9115	0.9030	0.9240	0.9126
朴素贝叶斯	0.7455	0.6659	0.9890	0.7956
人工神经网络	0.8695	0.9226	0.8460	0.8818
AdaBoost	0.9415	0.9335	0.9500	0.9412
Bagging	0.9230	0.9174	0.9310	0.9234
随机森林	0.9415	0.9355	0.9500	0.9421

三个集成学习模型的各性能指标均大于 0.91，与其他单一模型相比具有明显的优势。这得益于这些集成学习模型在模型构建中的集成机制，且对于含有 16 个输入变量的识别问题更加有效。

此外，同 URL 异常特征的结果一样，朴素贝叶斯表现是最差的，虽然查全率高达 0.9890，但 F 值、查准率和准确率却是最低的，分别为 0.7956、0.6659 和 0.7455，这意味着该方法将大多数正确网站识别为钓鱼网站，这与基于 URL 特征识别的结果恰恰相反。朴素贝叶斯是建立在特征变量相互独立的基础上的一种分类器[14]，很显然，本研究中收集的变量并非一定是独立的，这导致了该方法表现很差。

总的来看，基于网络特征数据进行识别的准确率较高，基本上可以正确地识别出钓鱼网站。

5.3 融合 URL 特征和网络评估数据的识别

(1) 特征选择

融合 7 个 URL 异常特征变量和 16 个网络评估数据特征，并采用 Boruta 进行特征变量的筛选，结果如图 4 所示。

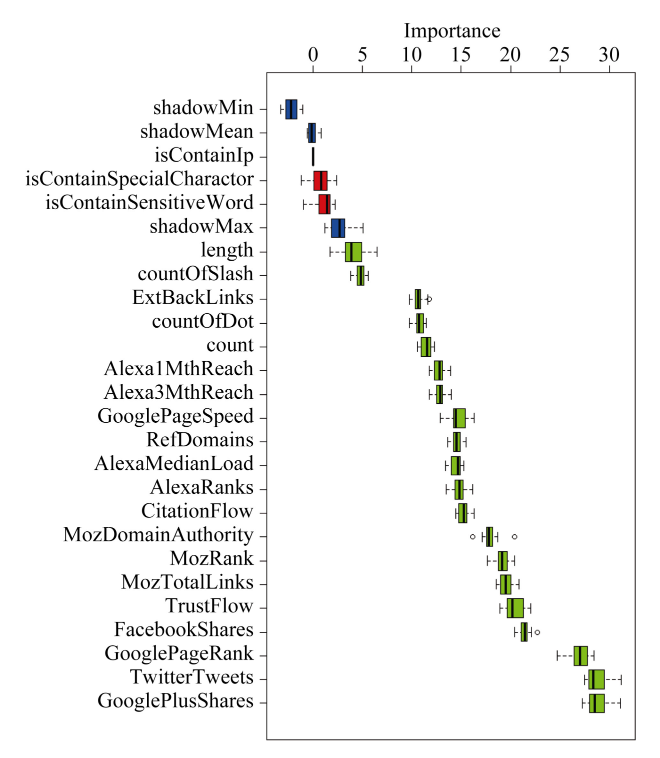


图 4 基于 Boruta 的 URL 和网络评估数据的特征选择

可知，3 个变量被拒绝，分别是 isContainIp(是否含有 IP)，isContainSensitiveWord(是否含有敏感词汇)以及 isContainSpecialCharactor(是否有特殊字符)，其余 20 个变量均被认为是重要的。

(2) 结果分析

采用融合 URL 特征与网络评估数据的 20 个变量构建钓鱼网站识别模型，实验结果如表 4 所示。同时，为了便于同基于 URL 异常特征的模型和基于网络评估数据的模型对比，图 5 给出了 8 种机器学习模型在采用不同变量特征时识别性能的对比结果。

表 4　8 种方法的评估结果

方法	准确率	查准率	查全率	F 值
决策树	0.8810	0.8576	0.9160	0.8845
SVM	0.9119	0.9280	0.9194	0.9185
K 近邻法	0.9200	0.9133	0.9300	0.9208
朴素贝叶斯	0.7690	0.6881	0.9880	0.8108
人工神经网络	0.8945	0.8879	0.8710	0.8776
AdaBoost	0.9415	0.9383	0.9430	0.9403
Bagging	0.9230	0.9174	0.9310	0.9234
随机森林	0.9435	0.9363	0.9530	0.9442

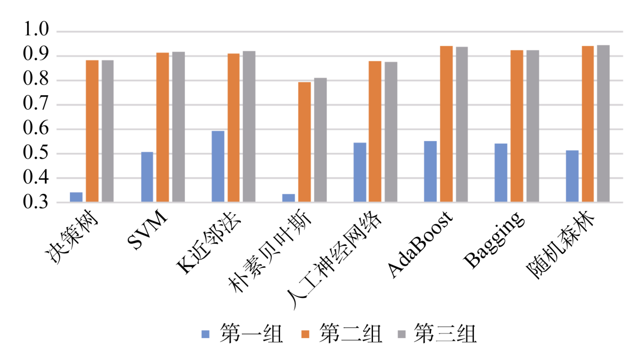


图 5　三组实验 F 值的对比

从表 4 可以看出，在融合两类数据特征的情况下，8 种机器学习识别模型中，朴素贝叶斯算法除查全率稍高之外，准确率、查准率和 F 值是最差的。如上文所述，这主要是由于变量间不一定相互独立所致。随机森林作为集成学习模型里的佼佼者，基于一定概率产生众多随机向量[15]，不仅可以有效构建多决策树以生成集成模型，还起到特征变量选择的作用。与已有研究中较好的表现一致[10, 12-13]，表 4 中的随机森林在准确率和 F 值上均是最好的。

对比图 5 中三组不同特征变量的实验结果可以看出，第二组和第三组的 F 值要远高于第一组，表明基于 URL 异常特征的识别模型并不能很好地进行钓鱼网站的识别，而基于网络评估数据以及融合两类特征的识别模型则能够较准确地识别钓鱼网站；同时，与只使用网络评估数据相比，融合两类特征的识别模型其结果准确率有一定的提高。这再次反映出网络评估数据在识别钓鱼网站中的有效性，也说明探索融合多种不同来源的特征变量以提高钓鱼网站识别性能是可行的。另外，相比于前 5 个单一模型，集成了多个弱学习器的集成模型取得了更高的 F 值，实现了更好地识别钓鱼网站的目的，再次表明了集成模型的高效与准确。

6　结　语

本文采用 8 种机器学习技术，对比研究了传统的基于 URL 异常特征的识别模型与最近的基于网络评估数据的识别模型在识别钓鱼网站问题上的性能，并融合两类特征变量，探究提高钓鱼网站识别准确性的可行性方案。实证结果表明，基于 URL 异常特征的识别模型不能很好地进行钓鱼网站的识别，而基于网络评估数据的识别模型具有较高的识别准确性，且融合两类特征向量的识别模型对钓鱼网站的识别准确性有一定提高。由于网络评估数据收集便捷，处理方式较为简单，因此在已有识别技术的基础上融合该类型特征变量是值得应用并推广的。

然而，在实际生活中，钓鱼网站与正常网站的比例是不均衡的，在之后的研究中将针对这一类别不均衡问题，研究更先进的机器学习技术与识别模型。此外，网站页面信息是识别钓鱼网站的另一重要数据，未来会尝试融合包括 URL 特征、页面信息、网络评估数据等更多不同来源的特征变量，以进一步提高钓鱼网站的识别准确性。

参考文献：

[1] Sheng S, Weidman B, Warner G, et al. An Empirical Analysis of Phishing Blacklists[C]//Proceedings of the 6th Conference on Email and Anti-Spam, California, USA.2009: 112-118.

[2] Zhang Y, Egelman S, Cranor L, et al. Phinding Phish: Evaluating Anti-phishing Tools[C]//Proceedings of the 14th Annual Network and Distributed System Security Symposium. 2007: 381-192.

[3] Blum A, Warden B, Solaria T, et al. Lexical Feature Based Phishing URL Detection Using Online Learning[C]//Proceedings of the ACM Workshop on Artificial Intelligence & Security. 2010: 54-60.

[4] 黄华军，钱亮，王耀钧. 基于异常特征的钓鱼网站 URL 检测技术[J]. 信息网络安全，2012 (1): 23-25. (Huang Huajun, Qian Liang, Wang Yaojun. Detection of Phishing URL Based on Abnormal Feature[J]. Netinfo Security, 2012(1): 23-25.)

[5] Ma J, Saul L K, Savage S, et al. Identifying Suspicious URLs: An Application of Large-scale Online Learning[C]//

Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009: 681-688.

[6] Ma J, Saul L K, Savage S, et al. Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2009: 1245-1254.

[7] 曾传璜, 李思强, 张小红. 基于 AdaCostBoost 算法的网络钓鱼检测[J]. 计算机系统应用, 2015, 24(9): 129-133. (Zeng Chuanhuang, Li Siqiang, Zhang Xiaohong. Phishing Detection System Based on AdaCostBoost Algorithm[J]. Computer Systems & Applications, 2015, 24(9): 129-133.)

[8] Thomas K, Grier C, Ma J, et a1. Design and Evaluation of a Real-time URL Spam Filtering Service[C]// Proceedings of the 2011 IEEE Symposium on Security and Privacy, Berkeley, California, USA. 2011: 376-382.

[9] 顾晓清, 王洪元, 倪彤光, 等. 基于贝叶斯和支持向量机的钓鱼网站检测方法[J]. 计算机工程与应用, 2015, 51(4): 87-90. (Gu Xiaoqing, Wang Hongyuan, Ni Tongguang, et al. Phishing Detection Approach Based on Naïve Bayes and Support Vector Machine[J]. Computer Engineering and Applications, 2015, 51(4): 87-90.)

[10] Hu Z, Chiong R, Pranata I, et al. Identifying Malicious Web Domains Using Machine Learning Techniques with Online Credibility and Performance Data[C]//Proceedings of the 2016 IEEE Congress on Evolutionary Computation (CEC), Vancouver, Canada. 2016: 5186-5194.

[11] Kursa M B, Rudnicki W R. Feature Selection with the Boruta Package[J]. Journal of Statistical Software, 2010, 36(11): 1-13.

[12] Freund Y, Schapire R E. A Decision-theoretic Generalization of On-line Learning and an Application to Boosting[J]. Journal of Computer and System Sciences, 1997, 55(1): 119-139.

[13] Lo S L, Chiong R, Cornforth D. Using Support Vector Machine Ensembles for Target Audience Classification on Twitter[J]. PLoS One, 2015, 10(3): 417-434.

[14] Bayes T, Price R, Canton J. An Essay Towards Solving a Problem in the Doctrine of Chances[J]. Reasonance, 2003, 8(4): 80-88.

[15] Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32.

作者贡献声明:

胡忠义, 王超群, 吴江: 提出研究思路, 设计研究方案;
王超群, 胡忠义: 数据处理, 负责实验, 论文起草;
胡忠义, 吴江, 王超群: 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: zhongyi.hu@whu.edu.cn。

[1] 胡忠义, 王超群, 吴江. Phishsites.csv. 钓鱼网站列表.

[2] 胡忠义, 王超群, 吴江. extractedFeatures.zip. 提取的特征向量.

收稿日期: 2017-04-10
收修改稿日期: 2017-05-29

Identifying Phishing Websites with Multiple Online Data Sources

Hu Zhongyi Wang Chaoqun Wu Jiang
(School of Information Management, Wuhan University, Wuhan 430072, China)
(The Center for Electronic Commerce Research and Development, Wuhan University, Wuhan 430072, China)

Abstract: [**Objective**] This study aims to identify phishing websites more effectively with the help of online evaluation data and URL abnormal features. [**Methods**] First, we used eight machine learning techniques to compare the performance of various online evaluation data and URL abnormal features in identifying phishing websites. Then, we proposed a new method to improve the accuracy of the identification procedures. [**Results**] We found that the evaluation data had better performance than abnormal features of URL. Combining the two data sets could improve the identification performance. [**Limitations**] We did not consider the difference between the numbers of phishing sites and the good ones. [**Conclusions**] Online evaluation data and URL abnormal features could help us identify phishing websites effectively, which indicates the direction of future studies.

Keywords: Data Mining Phishing Websites Identification Machine Learning

视听资料库系统 Avalon 媒体系统获得百万美元资助

由西北大学图书馆和印第安纳大学图书馆共同开发的视听资料库系统于近日获得 967 000 美元的联邦拨款，用于进一步提高档案机构管理和使大型视频和音频数字馆藏的能力。来自美国博物馆和图书馆服务研究所(Institute of Museum and Library Services, IMLS)的国家领导基金(编号: LG-70-17-0042-17)专门用于这个开源工具的功能开发和可持续性维护建设，该工具称为 Avalon 媒体系统。Avalon 旨在帮助机构为教师、学生和研究人员管理和提供音视频资料。

Avalon 的规划、设计和开发得到了来自 IMLS 和 Andrew W. Mellon 基金会的资助。西北大学和印第安纳大学主要负责该开源工具的软件开发，而各大机构则主要负责测试，以确保该系统满足研究、教学和文化遗产社区的需求。西北大学图书馆在两年半前实施了 Avalon，作为音视频资料存储库。目前，Avalon 在 6 个机构，包括西北大学、印第安纳大学、弗吉尼亚大学、华盛顿大学、阿尔伯塔大学和加尔文学院等 6 个机构全面实施，其他一些机构正处于不同的实施阶段。

该基金资助的为期两年的项目的 4 大目标包括:

(1) 将 Avalon 集成到 Samvera (Hydra)代码库中。Samvera 也称 Hydra，是一个大型数字存储库开源项目。Avalon 基于 Samvera 代码，并且已经在社区内进行了应用实施。目前的 Avalon 是独立于其他软件应用程序的，如果能集成到 Samvera 中，就可以从更广泛的 Samvera 社区中获得更好的支持和开发。

(2) 基于云的 Avalon 版本。基于云基础设施能使得 Avalon 平台易于安装和运行，进一步使供应商更容易提供 Avalon 服务，各种规模的机构也能更轻松地使用该工具。目前，没有专门 IT 人员的机构可能无法实施 Avalon，但是部署到端之后，实施将变得非常容易，机构可以像使用现成解决方案一样使用 Avalon。

(3) 改进媒体保存。虽然 Avalon 是一个用于长期存储数字文件的存储库，但它并不是一个数字保存系统。作为数字保存系统的存储需要长久的、健壮的数据存储保护，包括定期扫描损坏或丢失数据的“固定检查”。

(4) 实现标准化的交付格式。不同的数字平台相互交互时需要一个标准化的沟通桥梁。Avalon 技术团队将与相关组织一起合作，通过国际图像互操作性框架，制订一套音视频互操作性规范，并提供“示范实施”。

这 4 大目标的实现将使得 Avalon 成为获取音视频资料的最强大的解决方案，各种规模的档案机构将都能参与到保护和研究文化遗产中来。

(编译自: http://www.library.northwestern.edu/about/news/library-news/2017/imls-grant-award.html)

(本刊讯)